تلفیق مبانی فازی و یادگیری تقویتی در کنترل سیستم‌های دینامیکی*

مسعود گوهری منش[۱]   علی اکبر اکبری[۲]   محمد باقر نقیبی سیستانی[۳]

چکیده یادگیری تقویتی، روشی است که در آن عامل یا عاملان باتوجه به یک‌سری پاداش‌های مثبت و یا منفی، یک عمل بهینه را انجام می‌دهند. این روش، زمانی کارایی بسیار بالایی خواهد داشت که مدل سیستم به‌صورت طبیعی موجود نباشد و یا به‌دست آوردن آن موجب زحمت فراوان گردد. در این صورت می‌توان، آن را جایگزین مناسبی برای منطق‌های کنترلی دیگر دانست. یکی از معایب اساسی این روش، استفاده از عمل‌های گسسته در حین انجام آن می‌باشد. این در حالی است که خیلی از سیستم‌های دینامیکی با چنین رویکردی، عملکرد بهینه‌ای نخواهند داشت. برای جبران این نقیصه، رویکردهای متفاوتی از جمله تقریب مقادیر ظهور پیدا می‌کنند. در این مقاله از منطق فازی برای پیوسته کردن عمل‌های بهینه استفاده شده است. در این حالت، سیستم یادگیری تقویتی، قوانین بین کنترل‌کنندهٔ فازی را در جهت نیل به بهینه‌ترین عمل تنظیم می‌نماید و به این ترتیب می‌تواند عمل‌های پیوسته‌ای را تولید نماید. به این منظور مدل یک آونگ معکوس در سیم مکانیکس در نظر گرفته شده است که توسط کنترل‌کننده طراحی شده است و حرکت آن در دو حالت کنترل زاویهٔ آونگ و کنترل کامل آونگ و ارابه مورد بررسی قرار می‌گیرد. نتایج به‌دست آمده نشان می‌دهند، هوش مصنوعی به‌کار گرفته شده به‌جای انتخاب قوانین موجود، می‌تواند کارایی بالاتری در کنترل سیستم‌های دینامیکی داشته باشد.

واژه‌های کلیدی یادگیری تقویتی، منطق فازی، آونگ معکوس.

Combining the Principles of Fuzzy Logic and Reinforcement Learning for Control of Dynamic Systems

M. Goharimanesh   A.A. Akbari   M.B. Naghibi-Sistani

Abstract *Reinforcement learning is a method in which agent/agents obtain a positive or negative reward to do an efficient operation. In this way, the performance will be very suitable for the systems which are naturally complicated for deriving the differential equations. This can be a good alternative to other control areas. One of the main disadvantages of this method is considering the discrete actions during it. However, many of dynamical systems couldn't be optimized by this approach. To remedy this deficiency, different approaches have emerged, including approximate methods. In this paper, fuzzy logic is used to continually optimize the operations. In this case, the reinforcement learning method sets the fuzzy control rules which are the principles of optimal control. Two approaches, stabilizing the pendulum and both of pendulum and cart are considered to control the pole- cart problem in this paper. The results show that the applied artificial intelligence can be used as a proper solution for the taken policy.*

Key Words Reinforcement learning, Fuzzy logics, Inverse pendulum.

*تاریخ دریافت مقاله ۹۲/٦/۱۰ و تاریخ پذیرش آن ۹۳/۸/۵ می‌باشد.

(۱) دانشجوی دکتری مهندسی مکانیک، دانشگاه فردوسی مشهد

(۲) نویسندهٔ مسئول: دانشیار گروه مهندسی مکانیک، دانشگاه فردوسی مشهد. akbari@um.ac.ir

(۳) دانشیار گروه مهندسی برق، دانشگاه فردوسی مشهد

مقدمه

کنترل سیستم‌های دینامیکی، با در نظر گرفتن مدل سیستم، همواره مورد توجه قرار گرفته است. هنگامی که مدل، دست‌خوش تغییراتی چون عدم قطعیت گردد، نیز با ورود تکنیک‌هایی هم‌چون کنترل مقاوم این مسأله راه حل مناسبی پیدا می‌کند. اما چنان‌چه، مدل به‌طور شفاف مشهود نباشد و یا نتوان معادلات دینامیکی سیستم را استخراج نمود، می‌توان از کنترل بر پایهٔ یادگیری تقویتی استفاده کرد [1]. در این شیوه، کنترل‌کننده توسط یک نقاد بیرونی تنبیه و یا تشویق می‌گردد و براساس همان مواضع، می‌تواند خبرگی لازم را در جهت کنترل دینامیک موجود به‌دست آورد [2]. در این روش، دینامیک سیستم، محیط نام‌گذاری می‌گردد و خروجی‌های کنترل‌کننده به‌عنوان عامل یا عاملان می‌توانند کارایی لازم را به‌وجود بیاورند. انتخاب حالت‌ها، یکی از مهم‌ترین قسمت‌های این روش می‌باشد که بایستی توسط طراح، هوشمندانه صورت پذیرد. در واقع، حالات، همان بازخوردهایی هستند که در طی حل دینامیک سیستم به کنترل وارد می‌گردند. در یادگیری تقویتی، روش‌های مختلفی برای خبره کردن هوش سیستم وجود دارد. یکی از این روش‌ها، یادگیری کیو (Q-Learning) می‌باشد [3]. در این روش، حالت‌ها و عمل‌ها به‌صورت یک ماتریس قرار می‌گیرند. این جدول پس از گذشت تکرارهای انجام گرفته، به یک سیاست بهینه مبدل می‌شود که عامل می‌تواند با توجه به قرارگیری حالت فعلی، عمل بهینهٔ خود را تعیین نماید.

کاربردهای روش‌های یادگیری تقویتی طیف وسیعی را به خود اختصاص می‌دهند. در [4-6] از یادگیری تقویتی برای کنترل پرواز استفاده شده است. در [7,8] از این روش برای کنترل ربات بهره گرفته شده است. در [9] از یادگیری تقویتی برای حل مسائل بهینه‌سازی استفاده شده است. در [10] با استفاده از این روش، پارامترهای تنظیم‌کنندهٔ کنترل PID یک موتور خودرو مورد استفاده قرار گرفته است. در [11,12] کنترل واژگونی خودرو با استفاده از سیستم تعلیق نیمه‌فعال کنترلی طراحی شد که با رویکرد یادگیری تقویتی، ضرایب یک قانون کنترلی به‌دست می‌آمد. در [13] کنترل تعلیق یک خودروی سواری با یادگیری تقویتی مطرح شده است. در تعقیب مسیر و برنامه‌ریزی آن، مطالعاتی نیز صورت گرفته است که در اکثر موارد از مدل‌های روباتیکی و کوچک استفاده شده است [14]. در کنترل دینامیک خودرو در رفتار عرضی و کنترل واژگونی خودرو، کنترلی با استفاده از فازی و یادگیری تقویتی بر روی خودروهای سنگین صورت پذیرفت [15]. در کنترل حالت شارژ خودروهای هیبریدی، نیز الگوریتم‌های یادگیری تقویتی نیز توانسته‌اند به‌خوبی توانایی خود را نشان دهند [16]. در کنترل خودروهای بی‌سرنشین، مطالعات گوناگونی صورت گرفته است. در [17] کنترل یک هواپیمای مدل و در [18] کنترل هلیکوپتر برای مانورهای مختلف انجام گرفته است. کنترل دوچرخه به‌عنوان یک خودروی ناپایدار نیز مورد مطالعه قرار گرفته است [19]. پارک خودرو، حرکت رو به عقب خودروها، مخصوصاً خودروهای تجاری و چند مفصلی نیز چالشی بزرگ می‌باشد که راه حل مناسبی با استفاده از یادگیری تقویتی می‌توان برای آن یافت. اگرچه مطالعات صورت گرفته در این مورد خاص صرفاً بر روی سینماتیک حرکت است و به رفتار دینامیکی خودرو بهایی داده نشده است [20].

به‌دلیل رفتار غیرخطی مواد هوشمند که اخیراً نیز کاربرد بسزایی در علوم مختلف دارا می‌باشند، نمی‌توان مدل دقیقی را برای آنها متصور شد، لذا روش‌هایی که لزوماً مبتنی بر مدل نیستند و می‌توانند با استفاده از داده‌های محیطی کنترل این مواد را به عهده بگیرند بسیار مهم هستند [21,22]. در [23] کنترل یک مادهٔ هوشمند مغناطیسی با یادگیری تقویتی صورت گرفته است.

در [24] کنترل مفاصل بیومکانیکی مورد مطالعه قرار گرفته است. هم‌چنین در مورد ربات‌های موازی [25]، انسان‌نما [26] و ربات‌های صنعتی، مطالعات وسیعی صورت گرفته است [27]. در این زمینه نیز، اخیراً یک مطالعه مروری در [28] انجام شده است که به کاربرد الگوریتم‌های مبتنی بر یادگیری تقویتی در رباتیک می‌پردازد. در کنترل عملگرهای مکانیکی صنعتی، موارد [29,30] مطالعاتی را پیرامون کنترل حرکت مکانیزم بال- اسکرو (Ball-Screw) انجام داده‌اند. در زمینه تولید و فرآیندهای صنعتی، [31] مطالعه‌ای را بر روی کنترل چرخهٔ تولید توسط یادگیری تقویتی داشته است.

در مطالعه‌های بیولوژیکی نیز یادگیری تقویتی در کنترل بیماری ایدز [32]، تصفیهٔ آب [33,34] و کنترل سرطان [35] نیز کاربرد داشته است. در برخی مطالعات از یادگیری تقویتی به‌عنوان یک ابزار کمکی برای بهینه کردن کنترل‌کنندهٔ موجود گزارش شده است. به‌عنوان مثال انتخاب بهینهٔ پارامترها در یک کنترل‌کنندهٔ مد لغزشی توسط یادگیری تقویتی سبب کم شدن تصادم (Chattering) کنترل‌کننده شده است [36]. در [37] ارتباط میان یادگیری تقویتی و کنترل تطبیقی و پیاده‌سازی روش پیشنهادی را بر روی یک ربات انسان‌نما نشان می‌دهد. کنترل آونگ معکوس در اکثر مطالعات نیز به‌عنوان یک مثال کلاسیک از شروع ایده‌پردازی بر روی یادگیری تقویتی و مقایسهٔ الگوریتم‌های گوناگون مطالعه شده است.

یکی از محدودیت‌هایی که الگوریتم‌های یادگیری تقویتی دارند، استفادهٔ غیرپیوسته از عمل‌هایی است که در حین ساخت سیاست بهینه از سوی عامل تعیین می‌شود. گسستگی در عمل‌ها، نه‌تنها سیاست بهینه‌ای را در حد عالی به‌وجود نمی‌آورد بلکه ممکن است در خیلی از سیستم‌های دینامیکی حتی کاستی‌هایی اندک را نیز برطرف ننماید. کاستی صورت گرفته به آن دلیل است که اصل یادگیری تقویتی بر پایهٔ به‌روزرسانی هر حالتی از سیستم در هر مرحله می‌باشد. چنان‌چه در یک سیستم، تعدد حالات زیادتری اتفاق افتد، این امر سبب پایین آمدن کارایی الگوریتم یادگیری تقویتی می‌گردد. به همین دلیل از روش‌های تقریب به این منظور استفاده شده است. یکی از رویکردهایی که می‌توان در حل این معضل استفاده نمود، راهکاری است که توسط منطق فازی پی‌ریزی می‌شود [38,39]. از ۱۹۹٤ که برنجی [40] استفاده از فازی را به‌عنوان پیوسته کردن سیستم‌های دینامیکی معرفی نمود تا امروز، بیشتر رویکردهای مورد استفاده در کنترل که از یادگیری تقویتی استفاده می‌کنند، از فازی بهره گرفته‌اند.

ادامهٔ مقاله به این صورت تنظیم شده است که ابتدا استراتژی یادگیری تقویتی و روش‌های مبتنی بر آن مرور و بررسی خواهند شد. پس از آن، طرح مسأله‌ای کلاسیک انجام می‌گیرد و مدل آونگ معکوس با در نظر گرفتن یک شبیه‌ساز خارجی، مورد مطالعه قرار خواهد گرفت. پس از آن، کنترل براساس منطق فازی و رویکرد یادگیری تقویتی معرفی و انجام می‌گیرد. در نهایت، نتایج این روش نمایش داده خواهد شد.

استراتژی یادگیری تقویتی

روش‌های زیادی برای ایجاد یک سیاست بهینه ساخته و پرداخته شده‌اند. روش‌های برنامه‌ریزی پویا (Dynamic Programming)، روش مونت کارلو (Monte Carlo)، سارسا (SARSA) و یادگیری کیو از این دست می‌باشند [1]. استراتژی مورد استفاده در این مقاله، یادگیری کیو می‌باشد. رابطهٔ کیو، با استفاده از معادلهٔ (۱)، ارتباط منطقی میان حالت‌های شکل گرفته در مسأله و عمل‌های آن را ایجاد می‌نماید.

توصیف پارامترهای مرتبط به آن در ادامه بیان خواهد شد.

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a'} Q(s,a) - Q(s,a)\right] \qquad (۱)$$

در مسألهٔ حاضر که حالت‌ها، قوانین فی مابین منطق فازی می‌باشند، ممکن است برخی از آنها هیچ‌گاه مورد مطالعه قرار نگیرند و در جدول کیو، خالی گذارده شوند. هم‌چنین، عمل‌های مورد استفاده که در جدول کیو، مقادیر گسسته‌ای می‌باشند، پس از شکل‌گیری در منطق فازی به‌صورت پیوسته به سیستم القاء می‌گردند. حالت فعلی سیستم را با s و عمل حاضر را با a نشان می‌دهند. پس از آن‌که حالت سیستم به‌صورت بازخورد تحویل داده می‌شود، با بررسی به‌عمل آمده، پاداش او تنظیم می‌گردد. این مقدار با r مشخص می‌شود. هم‌چنین، عمل صورت گرفته، عامل را به حالت بعدی s' نیل می‌دهد. عبارت $\max_{a'} Q(s,a)$ به این معنا است که مقدار بیشینه حالت آتی عامل یا همان s' به‌ازای عمل‌های مختلف چه مقدار می‌باشد. هم‌چنین، نرخ یادگیری با α مشخص می‌شود. زمانی که نرخ یادگیری صفر در نظر گرفته شود، گویی عامل هیچ چیزی را فرا نمی‌گیرد و هر عملی را که انتخاب می‌کند بر پایهٔ نتایج همان لحظه می‌باشد. اگر این نرخ به یک نزدیک شود، به این معنا است که عامل تنها به یافته‌های پیشین خویش بها می‌دهد و هیچ نتیجهٔ فعلی را در روند تصمیم‌گیری قرار نداده است. به‌راحتی می‌توان مشاهده نمود که اگر α برابر با یک شود، $Q(s,a)$ از معادله حذف می‌گردد. در آونگ معکوس کنترل‌شده میزان نرخ یادگیری ۰/۱ در نظر گرفته شده است. نرخ کاهش با γ مشخص می‌شود. این عدد اگر صفر باشد، عامل تنها به پاداش فعلی خود بسنده می‌کند. گویی تنها، در این لحظه جلوی پای خود را می‌بیند و فرصت طلبی را پیشهٔ خود می‌نماید. در صورتی که اگر یک باشد، خود را برای یک پاداش در آینده کاری آماده می‌کند. این‌که این دو عدد چند باشند، تقریباً جواب مشخصی ندارد و برای هر مسأله می‌تواند، متفاوت به‌دست آید. در این مسأله، این مقدار ۰.۵ در نظر گرفته شده است. در این نوشتار برای مصالحهٔ بین اکتشاف (explore) و استخراج (exploit) از الگوریتم حریصانه استفاده شده است. انتخاب عملکرد حریصانه یک وسیلهٔ عمومی و مؤثر برای تعادل اکتشاف و بهره‌برداری در یادگیری تقویتی است. این بدین معنی است که احتمال انتخاب بدترین عمل به‌اندازهٔ انتخاب بهترین عمل بعدی است. در این الگوریتم، در هر مرحله، عامل برای انتخاب عمل‌های خود، تنها به این بسنده می‌کند که بیشترین عمل را انتخاب نماید نه بهترین آنها را. به این ترتیب او حریصانه به‌دنبال مقادیر بیشتری از عمل‌های پیش‌روی خود می‌باشد. برای این‌که این رفتار عامل کاملاً حریصانه نباشد، در طی انتخاب عمل برای او می‌توان قیدی را قرار داد، به‌طوری که اگر به‌طور تصادفی، یک مقدار بیشتر از مقدار قید عامل باشد، عمل پیش‌رو حریصانه انتخاب گردد و در غیر این صورت عامل یک عمل را به‌صورت تصادفی انتخاب نماید تا در بین عمل‌های موجود یک کاوش منطقی نیز داشته باشد. چنین الگوریتمی را اپسیلون-گریدی ($\varepsilon-greedy$) می‌نامند که در آن بازهٔ تعیین مقدار اپسیلون بین صفر و یک می‌باشد. در این مسأله، مقدار پارامتر این الگوریتم ۰.۰۰۱ در نظر گرفته شده است. تعیین پاداش می‌تواند به صور متفاوتی اتفاق بیفتد. مثال آن‌که می‌توان حالاتی که منجر به زاویه و سرعت زاویه‌ای کمتر از ۱ است را با پاداش ۱ و سایر حالات را با پاداش منفی ۱، در نظر گرفت. هم‌چنین می‌توان از فاصله اقلیدسی هرکدام از حالت‌ها با حالت تعادل آنها استفاده نمود و یک تابع پاداش هم‌چون تابع زیر در نظر گرفت.

$$R(\alpha,\omega) = \log(1/((\alpha-0)^2+(\omega-0)^2)) \qquad (۲)$$

چنین تابعی سبب می‌شود که بـه حـالات نزدیـک تعادل، پاداش فوق‌العاده مثبتی تعلق بگیرد. در حالی که هر چقدر از آن حالات فاصله گرفته شود، پاداش، رنگ منفی‌تری به خود می‌گیرد. این خاصیت سبب می‌شـود یک پیوستگی در تشویق و تنبیه صورت پذیرد.

در مطالعهٔ حاضر، دو حالت مختلف بـرای کنتـرل آونگ توسط یادگیری تقـویتی و فـازی اسـتفاده شـده است. در مورد اول، تنها زاویه و سرعت زاویه‌ای آونگ کنترل‌شده است. در این حالـت، تـابع پـاداش بـه ایـن صورت تعریف می‌شود که برای زوایـای کمتـر از ۰/۵ درجه و سرعت زاویـه‌ای کمتـر از ۱ درجـه بـر ثانیـه، پاداش ۱ و برای حالتی کـه زاویـه و سـرعت زاویـه‌ای آونگ به‌ترتیب بیشتر از ۵ درجه و ۵۰ درجـه بـر ثانیـه باشد، مقدار منفی ۱ را لحاظ می‌کنـد. در سـایر مـوارد، پاداشی برابر صفر به عامل داده می‌شود.

در نگاه دوم کنترلی که مبتنی بر کنترل تمام آونـگ می‌باشد، چنان‌چـه بـه‌عـلاوهٔ زاویـه آونـگ و سـرعت زاویه‌ای آن در حالت قبل، موقعیت ارابه و سـرعت آن به‌ترتیب کمتر از ۰/۱ متـر و ۰/۱ متـر بـر ثانیـه باشـند، پاداشی برابر با ۱ و چنان‌چـه ایـن دو مـورد بـه‌ترتیـب بیشتر از ۰/۳ متر و ۰/۵ متر بر ثانیه باشند، پاداشی برابر با منفی یک به آن داده خواهد شد.

در مثال حاضر، آونگ در محیط سیمولینک ساخته می‌شود و سپس توسط سیگنال‌های بازخوردی، در هـر لحظه مورد ارزیابی قرار می‌گیرد و عمـل بهینـه در هـر زمان تعیین می‌شود.

مدل‌سازی و شبیه‌سازی آونگ معکوس

روش‌های زیادی برای مدل‌سازی و شبیه‌سـازی آونـگ معکوس وجود دارند. در [39] معادلات آونگ معکوس به‌صورت زیر آورده شده است.

$$\ddot{\theta}=\frac{g\sin(\theta)+\cos(\theta)\left(-f-ml\ddot{\theta}^{2}\sin(\theta)+\mu_c\,\mathrm{sgn}(\dot{x})\right)/(m_c+m)-\mu_p\dot{\theta}/ml}{l\left(4/3-\left(m\cos^{2}(\theta)\right)/\left(m_c+m\right)\right)}$$

$$\ddot{x}=\frac{f+ml\left(\ddot{\theta}^{2}\sin(\theta)-\ddot{\theta}\cos(\theta)\right)-\mu_c\,\mathrm{sgn}(\dot{x})}{m_c+m}$$

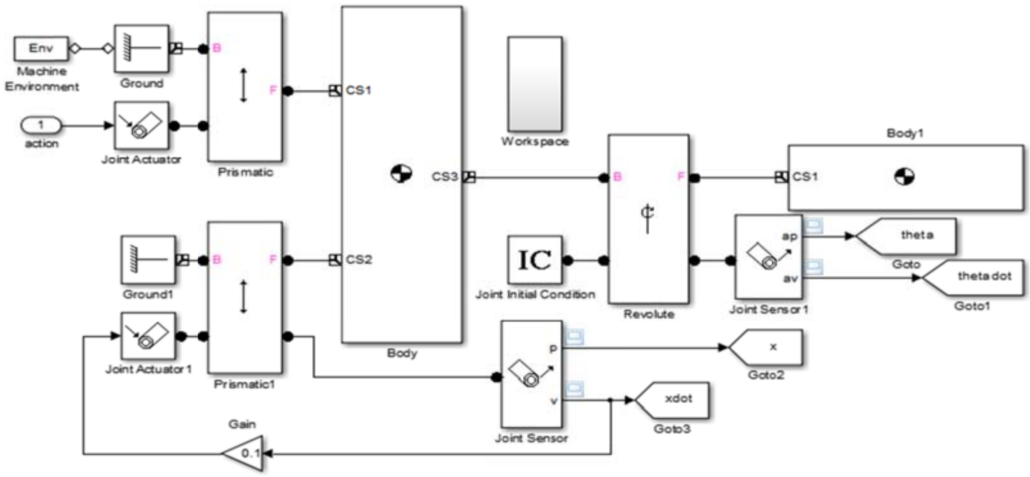
(۳)

در جـدول (۱) توصـیف هـر پـارامتر آورده شـده است.

جدول ۱ پارامترهای موجود در معادلات آونگ معکوس

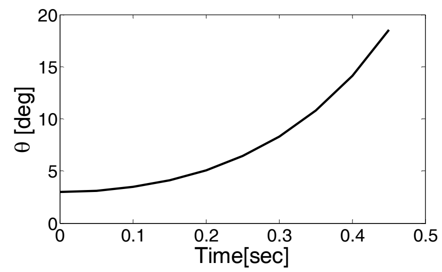
زاویهٔ آونگ، نسبت به محور عمود برحسب درجه (مثبت در سمت چپ. منفی در سمت راست)	θ
سرعت زاویه‌ای آونگ برحسب درجه بر ثانیه	$\dot{\theta}$
شتاب زاویه‌ای آونگ برحسب درجه بر ثانیه	$\ddot{\theta}$
شتاب گرانشی زمین برحسب متر بر مجذور ثانیه	g
نیروی وارد بر ارابه برحسب نیوتن	f
جرم ارابه برحسب کیلوگرم	m
جرم آونگ برحسب کیلوگرم	m_c
نصف طول آونگ برحسب متر	l
ضریب اصطکاک ارابه بر روی محور حرکتی	μ_c
ضریب اصطکاک آونگ بر روی ارابه	μ_p
موقعیت حرکتی مرکز ثقل ارابه برحسب متر	x
سرعت خطی مرکز ثقل ارابه برحسب متر بر ثانیه	$\dot{x}$
شتاب خطی مرکز ثقل ارابه برحسب متر بر مجذور ثانیه	$\ddot{x}$

یکی از مشکلاتی کـه اکثـر معـادلات آورده شـده دارند، در نظر نگرفتن حالـت واقعـی آونـگ معکـوس به‌صورت یک جرم صلب مـی‌باشـد. چـرا کـه در ایـن حالت هم ارابه و هم آونـگ معکـوس، ممـان اینرسـی مورد نظر خود را به‌دلیل شکل، ظاهر و توزیع وزن دارا هستند که اکثـر ایـن مـوارد در ایـن معـادلات بررسـی نمی‌شوند. این در حالی است کـه اسـتفاده از یـادگیری تقویتی به همین دلیل می‌باشد که سیستم از نگـاه نقـاد کنترلی فاقد شناخت می‌باشد و تنها به صـرف دریافـت بازخوردهـای لازم از محـیط، عمـل کنترلـی انجـام می‌پذیرد.

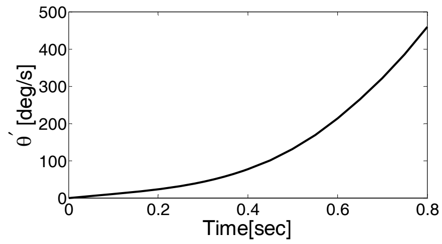


شکل ۱ مدل کامل آونگ معکوس با در نظر گرفتن نیروی اصطکاک در سیم مکانیکس

در این مقاله، مدل حاضر در فضای سیم مکانیکس (Simmechanics) در متلب شبیه‌سازی می‌گردد. حلگرهایی که در این روش برای حل سینماتیک و دینامیک ربات استفاده می‌شوند براساس روش‌های تکرارپذیری انرژی حل می‌شوند. در این برنامه، آونگ به‌صورت تقریباً واقعی موجودیت پیدا می‌کند و با ارتباطی که میان سیستم کنترلی و نقاد مربوط صورت می‌پذیرد، عمل کنترل انجام خواهد گرفت. در تصویر (۱) مدل آونگ معکوس با در نظر گرفتن تمامی اجزای آن در فضای سیمولینک و با استفاده از سیم مکانیکس ترسیم شده است. در این نوشتار جرم ارابه و آونگ به‌ترتیب ۰/۵ و ۰/۲ کیلوگرم گذاشته شده است. هم‌چنین ممان اینرسی حول محور دوران، برای ارابه و آونگ، به‌ترتیب ۱ و ۰/۰۰۶ کیلوگرم مترمربع در نظر گرفته شده است. اصطکاک ویسکوز نیز میان زمین و ارابه برابر با ۰/۱ قرار داده شده است. با در نظر گرفتن مقادیر اولیهٔ ۳ درجه برای زاویه و سرعت زاویه‌ای صفر درجه بر ثانیه برای آونگ، می‌توان شبیه‌سازی این مدل را در شکل‌های (۳ و ۲) که به‌ترتیب بیانگر زاویه و سرعت زاویه‌ای آونگ کنترل نشده می‌باشند، مشاهده نمود.



شکل ۲ زاویه آونگ معکوس کنترل‌نشده بر حسب زمان در طی ۰.۵ ثانیه



شکل ۳ سرعت زاویه‌ای آونگ معکوس کنترل‌نشده بر حسب زمان در طی ۰.۵ ثانیه

کنترل آونگ معکوس

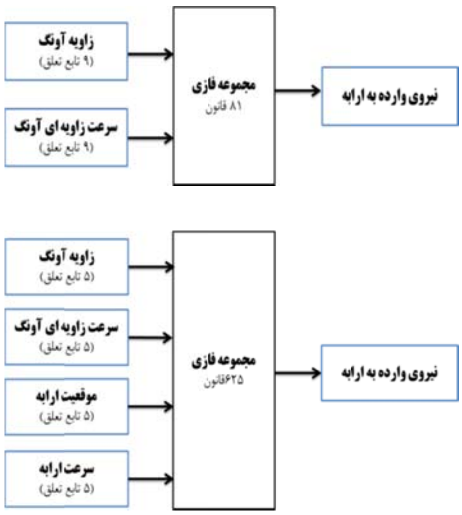
نظریهٔ مجموعه‌های فازی در سال ۱۹٦۵ مطرح شد [41]. بزرگ‌ترین مشکلی که سر راه سیستم‌های فازی می‌باشد ایجاد قوانین لازم برای ارتباط بین ورودی‌ها و خروجی‌ها است که بایستی توسط یک فرد خبره انجام پذیرد. در این قسمت، یادگیری تقویتی، قوانین بهینه را

تولید و در مدار اجرای کنترل‌کنندهٔ فازی قرار خواهد داد. هر سیگنالی که از سیستم دریافت می‌گردد در واقع یک ورودی برای سیستم فازی محسوب می‌شود، پس از تعیین مقدار تابع عضویت برای هر کدام از ورودی‌ها، قانون متناظر با آنها استخراج و متناظر با آن قانون، حالت مشخصه در جدول کیو تعیین می‌شود و براساس تکنیک‌هایی نظیر اپسیلون گریدی (epsilon greedy) یا سافت مکس (softmax)، عمل مربوط تعیین و در خروجی فازی قرار می‌گیرد. سیگنال‌های بازگشتی به عامل در قسمت اول، تنها زاویهٔ آونگ و سرعت زاویه‌ای آن برحسب درجه و درجه بر ثانیه می‌باشند و در قسمت دوم دو مورد موقعیت حرکت ارابه و سرعت آن نیز به‌منظور کنترل دقیق‌تر، اضافه می‌گردند. در شبیه‌سازی اول این مقاله ۹ تابع تعلق برای ورودی اول و ۹ تابع برای ورودی دوم در نظر گرفته می‌شود. در مجموع ۸۱ حالت و یا قانون برای سیستم فازی می‌توان متصور شد. در قسمت دوم برای کاهش میزان محاسبات برای هر ورودی، ۵ تابع تعلق در نظر گرفته شده است که در مجموع ٦۲۵ قانون بایستی احراز شوند. خروجی هر قانون به‌صورت یک عمل از یک مجموعه عمل‌های گسسته تعیین می‌شود که برای مسألهٔ حاضر می‌توان مجموعه اعداد شکل (٤) را در نظر گرفت.

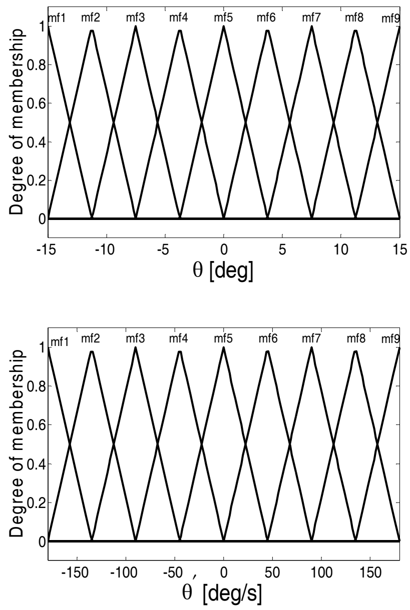
$$A = \{-10, -8, -6, -4, -2, 0, 2, 4, 6, 8, 10\} \qquad (٤)$$

بر این اساس جدول کیو، با تعداد سطر متناظر با تعداد قوانین و ۱۱ ستون برای ارزش‌گذاری هر سلول در جهت تعیین سیاست بهینه استفاده می‌شود. شکل (٤) این ارتباط را در یک کنترل‌کنندهٔ فازی نشان می‌دهد. شکل اول، مربوط به مجموعه‌ای است که در آن تنها حرکت آونگ کنترل می‌شود و مجموعهٔ دوم حالتی است که کنترل کامل آونگ و ارابه را با یکدیگر انجام می‌دهد.

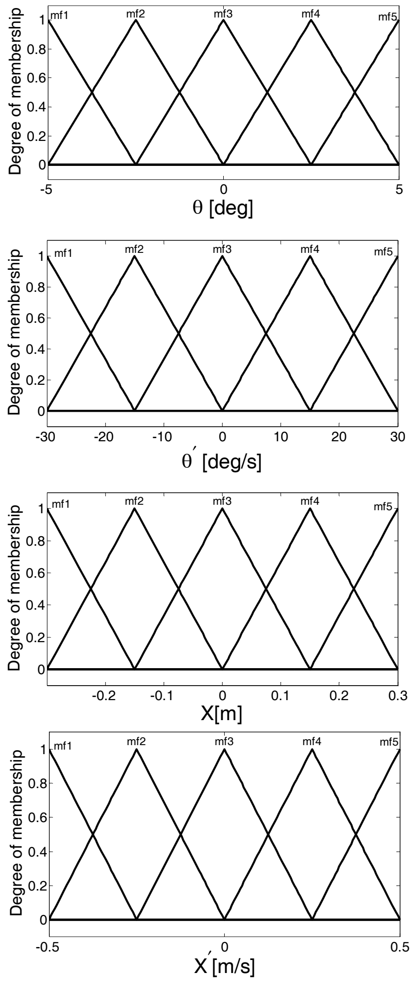
در شکل‌های (٦ و ۵) توابع عضویت هر کدام از ورودی‌های سیستم فازی برای هر دو مجموعهٔ فازی به نمایش گذاشته شده است. در هر کدام از این دو شکل، ۹ تابع تعلق مثلثی شکل استفاده شده است که درجه عضویت زاویه و سرعت زاویه ای برای هر یک بین صفر تا ۱ مشخص شده‌اند.



شکل ٤ مجموعهٔ فازی



شکل ۵ توابع تعلق ۹ گانه مربوط به حالت زاویه و سرعت زاویه‌ای آونگ

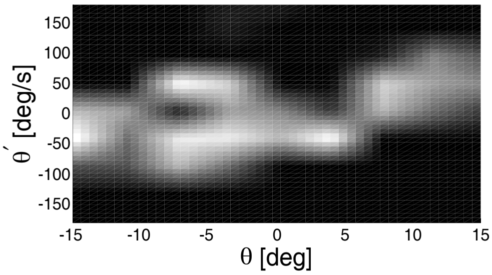


شکل ٦ توابع تعلق ٥ گانه مربوط به ٤ حالت زاویه و سرعت آونگ و موقعیت و سرعت ارابه

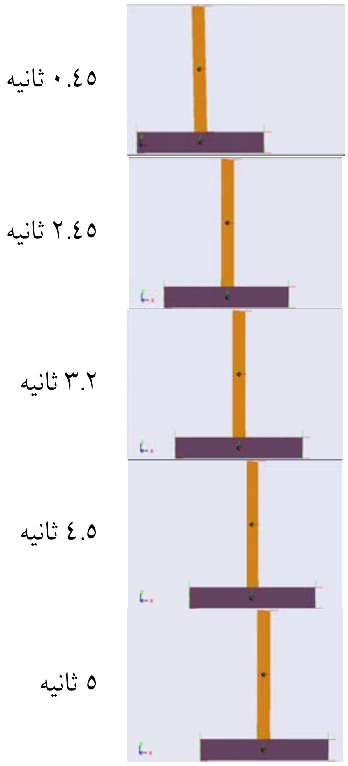
کنترل حرکت عمودی آونگ

در شکل (٧) سیاست بهینهٔ مورد نظر نشان داده شده است. در واقع این شکل نشان می‌دهد که برای هر زاویه و هر سرعت زاویه‌ای چه مقدار نیرو به‌عنوان عمل بایستی ایجاد شود تا پایداری آونگ محرز گردد. قسمت‌های روشن‌تر به‌معنای مقادیر بیشتر و قسمت‌های تاریک‌تر به‌معنای نیروی کمتر می‌باشند.

شکل (٨) دستاورد این کنترل‌کننده را در یک اجرا در زمان‌های مختلف نشان می‌دهد و مشخص می‌کند که سیستم به پایداری لازم دسترسی پیدا کرده است.



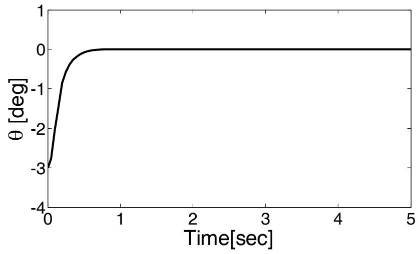
شکل ٧ سیاست بهینهٔ به‌دست آمده از یادگیری آونگ براساس دو حالت زاویه و سرعت زاویه‌ای آونگ



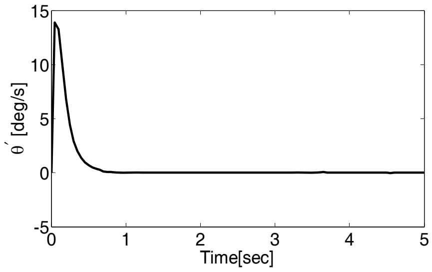
شکل ٨ موقعیت آونگ معکوس کنترل‌شده در زمان‌های متوالی و متفاوت

همان‌طور که در شکل (٨) نشان داده شده است، شبیه‌سازی اول مقاله با در نظر گرفتن زاویه و حرکت زاویه‌ای آونگ تنها می‌تواند حالت عمودی آن را کنترل نماید و همان‌طور که در نتایج مربوط به موقعیت ارابه نشان داده شده است، حرکت ارابه هیچ کنترلی بر روی آن صورت نگرفته است. در شکل‌های (١٢-٩) به‌ترتیب زاویه، سرعت زاویه‌ای آونگ، موقعیت ارابه و

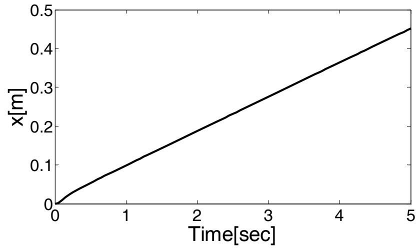
سرعت آن بر حسب زمـان در طـی ۵ ثانیـه نشـان داده شده‌اند.



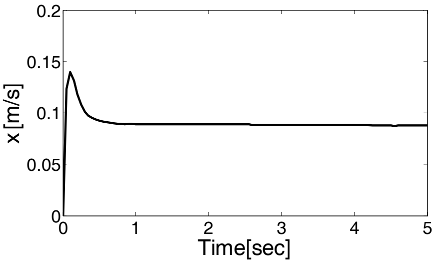
شکل ۹ زاویهٔ کنترل‌شدهٔ آونگ بر حسب زمان



شکل ۱۰ سرعت زاویه‌ای کنترل‌شدهٔ آونگ بر حسب زمان

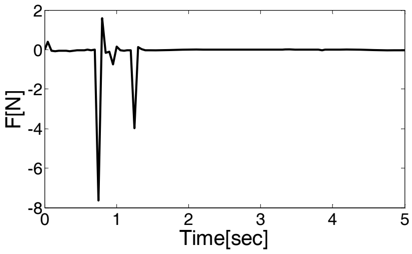


شکل ۱۱ موقعیت کنترل‌نشدهٔ ارابه بر حسب زمان

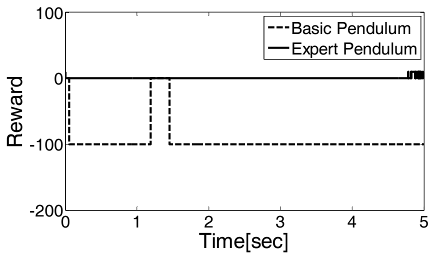


شکل ۱۲ سرعت کنترل‌نشدهٔ ارابه بر حسب زمان

در تصویر (۱۳) میزان نیرویـی کـه جهـت کنتـرل آونگ معکوس توسط سیستم فازی-یـادگیری تقـویتی تأمین شده است نمـایش داده مـی‌شـود. ایـن تصـویر، پیوسته بودن نتـایج یـادگیری تقـویتی را بـا اسـتفاده از سیستم فازی نمایش می‌دهد.



شکل ۱۳ نیروی کنترلی آونگ معکوس کنترل‌شده بر حسب زمان



شکل ١٤ مقدار پاداش‌های داده‌شده بر حسب زمان برای آونگ مبتدی و خبره

شکل (١٤)، مقدار پاداش‌های کسب‌شده براسـاس مدل پیشنهادی تابع پاداش در بخش‌هـای پیشـین را در یک اجرا برای یک آونگ خبره و یک آونگ مبتدی بـه نمایش می‌گذارد. همان‌طور که نشان داده شـده اسـت، مقدار پاداش‌های کسب‌شده برای آونگ مبتـدی رو بـه کاهش می‌باشد. این در حالی است که آونگ خبـره در طی زمان به‌دلیل کنترل زاویه و سرعت زاویـه‌ای خـود می‌تواند پاداش‌های بیشتری را کسب نماید.

کنترل کامل آونگ و ارابه

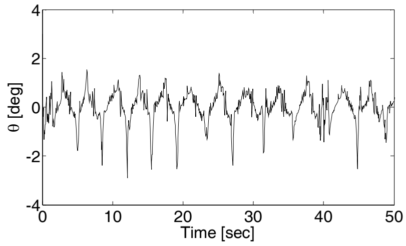
در قسمت قبل کنترل آونگ تنها به‌منظور حفـظ حالـت عمودی آونگ بود و حرکـت رو بـه جلـو و یـا رو بـه عقب ارابه در نظر گرفته نشده است. در این بخش، دو حالت جدید یعنـی موقعیـت و سـرعت ارابـه نیـز بـه حالت‌های پیشین اضافه و کنترل آونگ به ایـن ترتیـب

انجام می‌گیرد. نتایج به‌دست آمده برای سیستم آونگ در شکل‌های (۲۵-۱۵) نشان داده شده است. برای مقایسهٔ کنترل طراحی شده با یک کنترل‌کننده دیگر، منطق فازی بار دیگر استفاده می‌شود و این بار به‌منظور پیدا کردن تعداد بالای قوانین (٦٢٥) از ١٦ قانون که توسط یک فرد خبره مشخص شده است، استفاده می‌گردد. در این حالت برای هر کدام از چهار حالت، دو تابع تعلق بیشتر در نظر گرفته نمی‌شود. قوانین مورد استفاده به‌صورت جدول (۲) می‌باشند. در این جدول، سطرها مختص موقعیت و سرعت ارابه و ستون‌ها بیانگر زاویه و سرعت زاویه‌ای آونگ می‌باشند که حرف N به‌معنای کم‌ترین مقدار در تابع عضویت آن ورودی و P به‌معنای بیشترین مقدار آن می‌باشد. در داخل جدول نیز حروف N، P و Z به‌معنای کم‌ترین، بیشترین و مقدار صفر نیروی واردشده می‌باشد که از مجموعه اعداد رابطهٔ (٤) نتیجه می‌شوند.

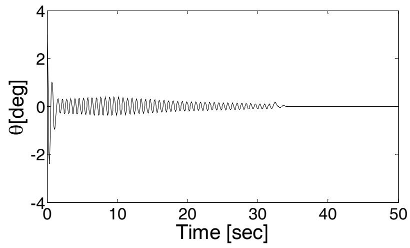
جدول ۲ قواعد فازی نگاشته شده توسط یک خبرهٔ انسانی

قوانین	پاندول	سرعت-زاویه			
سرعت-موقعیت	ارابه	NN	NP	PN	PP
	NN	P	P	N	N
	NP	P	Z	Z	N
	PN	P	Z	Z	N
	PP	P	P	N	N

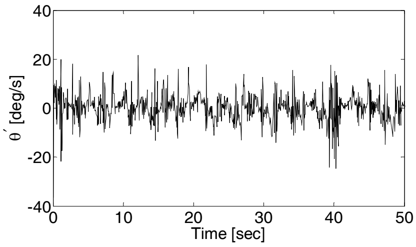
در نمودارهای (٢٤-١٥) زاویهٔ آونگ، سرعت زاویه‌ای آونگ، موقعیت ارابه، سرعت ارابه و نیروی اعمال‌شده توسط کنترل‌کنندهٔ ارابه برای هر دو الگوریتم فازی پیشنهادی در حالت اخیر به نمایش گذاشته شده است. در مجموعهٔ فازی اول، قوانین توسط یادگیری تقویتی به‌دست آمده‌اند و در دومین مجموعه، قوانین طبق جدول (۲) گذاشته شده‌اند. همان‌طور که نمودارهای زاویه و سرعت زاویه‌ای آونگ نمایش می‌دهند، کنترل زاویهٔ آونگ در هر دو مجموعه سبب پایداری حرکت آونگ شده است اما، طبق آن‌چه که نمودار موقعیت ارابه نشان می‌دهد، یادگیری تقویتی در این مورد با خبرگی بیشتر عمل کرده است و موقعیت و سرعت حرکت ارابه را نیز به کنترل در آورده است. این واقعیت در تصویر (۲۵) که از لحظه نخست حرکت آونگ و به فواصل ۱۰ ثانیه گرفته شده است، این واقعیت را به‌خوبی نشان می‌دهد.



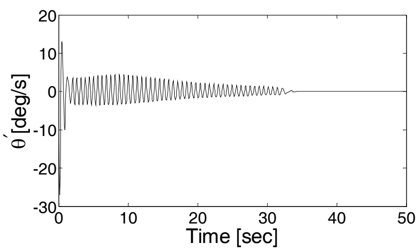
شکل ۱۵ زاویهٔ کنترل شدهٔ آونگ بر حسب زمان توسط فازی و یادگیری تقویتی



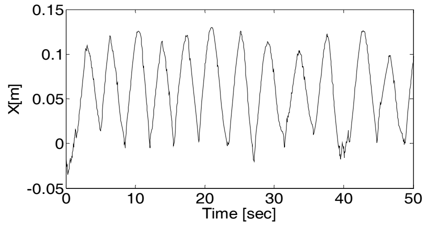
شکل ١٦ زاویهٔ کنترل شدهٔ آونگ بر حسب زمان توسط کنترل‌کنندهٔ فازی



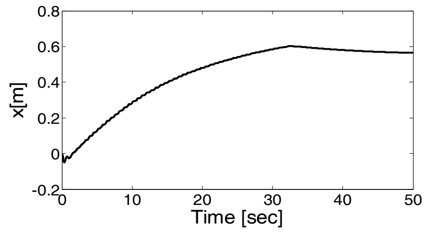
شکل ۱۷ سرعت زاویه‌ای کنترل‌شدهٔ آونگ بر حسب زمان توسط فازی و یادگیری تقویتی



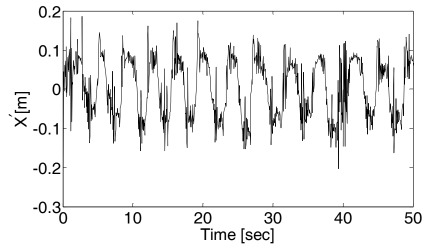
شکل ۱۸ سرعت زاویه‌ای کنترل‌شدهٔ آونگ بر حسب زمان توسط کنترل‌کنندهٔ فازی



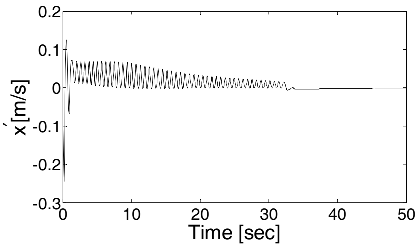
شکل ۱۹ موقعیت کنترل‌شدهٔ ارابه بر حسب زمان توسط فازی و یادگیری تقویتی



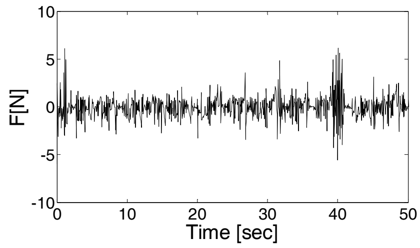
شکل ۲۰ موقعیت کنترل‌شدهٔ ارابه بر حسب زمان توسط کنترل‌کننده فازی



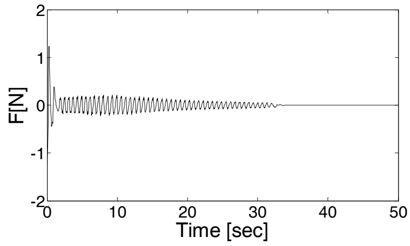
شکل ۲۱ سرعت کنترل‌شدهٔ ارابه بر حسب زمان توسط فازی و یادگیری تقویتی



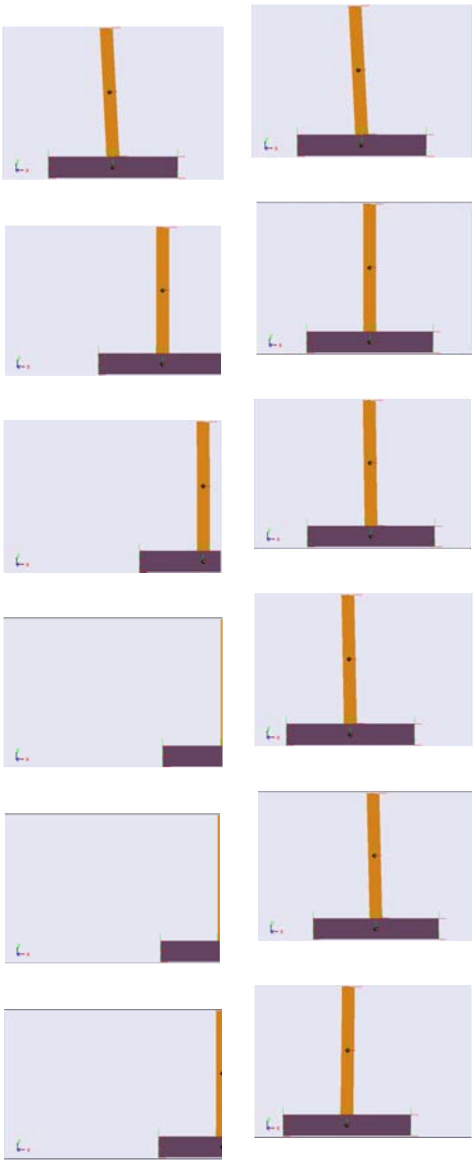
شکل ۲۲ سرعت کنترل‌شدهٔ ارابه بر حسب زمان توسط کنترل‌کننده فازی



شکل ۲۳ نیروی کنترلی اعمال شده به ارابه بر حسب زمان توسط فازی و یادگیری تقویتی



شکل ۲٤ نیروی کنترلی اعمال شده به ارابه توسط کنترل‌کنندهٔ فازی



الگوریتم فازی مبتنی بر یادگیری تقویتی

الگوریتم فازی مبتنی بر قوانین انسانی

شکل ۲۵ موقعیت آونگ معکوس کنترل‌شده در زمان‌های متوالی ۱۰ ثانیه در حالت کنترل کامل آونگ و ارابه

گرچه، با بیشتر کردن تعداد قوانین جدول (۲) می‌توان به کنترل بهتری دست پیدا کرد، اما بایستی توجه داشت که افزایش تعداد قوانین، پیدا کردن آن را نیز به چالش می‌کشد و زمان زیادی را برای یک شخص به خود اختصاص می‌دهد و این در حالی است که چنان‌چه مدل آونگ دستخوش تغییرات شود، کنترل‌کنندهٔ موجود نیز بایستی با یک سری قوانین تازه به کار خویش ادامه دهد در حالی که با استفاده از یادگیری تقویتی می‌توان قوانین را با تعدد بالای خود و هرگونه تغییر در مدل دینامیکی به‌وجود آورد.

نتیجه‌گیری

در این مقاله رویکرد یادگیری تقویتی برای بهینه کردن قوانین فازی در کنترل یک سیستم دینامیکی مورد توجه قرار گرفت. یادگیری تقویتی برای سیستم‌هایی که معادلات فیزیکی آنها به‌صورت کاملاً واضح مشخص نمی‌باشند، بیشتر مورد توجه قرار می‌گیرد. به همین دلیل در این مقاله از یک مدل ساخته شده در سیم مکانیکس که آونگ معکوس را شبیه‌سازی می‌نماید استفاده شد و توسط موارد به‌کار گرفته شده در اجرای یادگیری تقویتی، کنترل‌کننده فازی در طی آن خبره گردید. نتایج به عمل آمده با توجه به حالات مختلف استفاده از کنترل‌کنندهٔ طراحی‌شده و هم‌چنین در مقایسه با کنترل‌کننده‌های دیگر نشان می‌دهند که کنترل‌کنندهٔ فعلی از هوش کافی برای کنترل محیط ساخته شده برخوردار می‌باشد. چنین رویکردی می‌تواند برای هر سیستم دینامیکی مجهولی با بالاترین درجهٔ خبرگی مورد استفاده قرار گیرد.

مراجع

1. Sutton, R.S. and Barto, A.G., "*Reinforcement learning: An introduction"*, Vol. 1, Cambridge Univ Press, (1998).
2. Kaelbling, L.P., Littman, M.L. and Moore, A.W. "Reinforcement learning: A survey", *Journal of Artificial Intelligence Research,* Vol. 4, pp. 237-285, (1996).
3. Watkins, C.J. and Dayan, P., "Q-learning", *Machine learning,* Vol. 8, pp. 279-292, (1992).
4. Berenji, H. Lea, R. Jani, Y., Khedkar, P., Malkani, A. and Hoblit, J., "Space shuttle attitude control by reinforcement learning and fuzzy logic", in *Fuzzy Systems, Second IEEE International Conference on* 1993, pp. 1396-1401, (1993).
5. Graepel, T. Herbrich, R. and Gold, J., "Learning to fight", in *Proceedings of the International Conference on Computer Games: Artificial Intelligence, Design and Education*, pp. 193-200, (2004).
6. Ng, A.Y., Coates, A. Diel, M., Ganapathi, V., Schulte, J., Tse, B., Berger, E. and Liang, E., "Autonomous inverted helicopter flight via reinforcement learning", *Experimental Robotics IX*, ed: Springer, pp. 363-372, (2006).
7. Lin, C-.K. "A reinforcement learning adaptive fuzzy controller for robots," *Fuzzy Sets and Systems,* Vol. 137, pp. 339-352, (2003).
8. Yung, N.H. and Ye, C. "An intelligent mobile vehicle navigator based on fuzzy logic and reinforcement learning", *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on,* Vol. 29, pp. 314-321, (1999).
9. Barto, A. and Crites, R.,"Improving elevator performance using reinforcement learning", *Advances in neural information processing systems,* Vol. 8, pp. 1017-1023, (1996).
10. Howell, M. and Best, M.C., "On-line PID tuning for engine idle-speed control using continuous action reinforcement learning automata", *Control Engineering Practice,* Vol. 8, pp. 147-154, (2000).
11. Frost, G., Howell, M., Gordon, T. and Wu, Q., "Dynamic vehicle roll control using reinforcement

learning", (1996).

12. Howell, M.N. Frost, G.P. Gordon, T.J. and Wu, Q.H., "Continuous action reinforcement learning applied to vehicle suspension control", *Mechatronics,* Vol. 7, pp. 263-276, (1997).
13. Bucak, İ. and Öz, H., "Vibration control of a nonlinear quarter-car active suspension system by reinforcement learning", *International Journal of Systems Science,* Vol. 43, pp. 1177-1190, (2012).
14. Lauer, M., "A case study on learning a steering controller from scratch with reinforcement learning," in *Intelligent Vehicles Symposium (IV), 2011 IEEE*, pp. 260-265, (2011).
15. Akbari, A.A. and Goharimanesh, M., "Yaw Moment Control Using Fuzzy Reinforcemnt Learning", *Advanced Vehicle Control (AVEC14)*, (2014).
16. Vogel, A ., Ramachandran, D., Gupta, R. and Raux, A., "Improving Hybrid Vehicle Fuel Efficiency Using Inverse Reinforcement Learning", *AAAI*, (2012).
17. Woodbury, T., Dunn, C. and Valasek, J., "Autonomous Soaring Using Reinforcement Learning for Trajectory Generation", (2014).
18. Ng, A.Y., Kim, H.J. Jordan, M.I. Sastry, S. and Ballianda, S., "Autonomous Helicopter Flight via Reinforcement Learning", *NIPS*, (2003).
19. Cam, B. Dembia, C. and Israeli, J., "Reinforcement learning for bicycle control", (2013).
20. Yamashita, S., Horiuchi, T. and Kato, S., "A study on skill acquisition in trailer-truck steering problem by reinforcement learning", *SICE 2002. Proceedings of the 41st SICE Annual Conference*, pp. 810-812, (2002).
21. Kirkpatrick, K. and Valasek, J., "Reinforcement learning for characterizing hysteresis behavior of shape memory alloys", *Journal of Aerospace Computing, Information, and Communication,* Vol. 6, pp. 227-238, (2009).
22. Kirkpatrick, K. and Valasek, J., "Active length control of shape memory alloy wires using reinforcement learning", *Journal of Intelligent Material Systems and Structures,* Vol. 22, pp. 1595-1604, (2011).
23. Zhou, M., Hu, B., Gao, W. and Wang, J., "Reinforcement Learning Fuzzy Neural Network Control for Magnetic Shape Memory Alloy Actuator," *International Journal of Control & Automation,* Vol. 7, No. 6, pp. 109-122, (2014).
24. Uragami, D., Takahashi, T. and Matsuo, Y., "Cognitively inspired reinforcement learning architecture and its application to giant-swing motion control", *Biosystems,* Vol. 116, pp. 1-9, (2014).
25. Shahriari, M. and Khayyat, A.A., "Gait analysis of a six-legged walking robot using fuzzy reward reinforcement learning", *Fuzzy Systems (IFSC), 13th Iranian Conference on*, pp. 1-4, (2013)
26. Navarro-Guerrero, N., Weber, C., Schroeter, P. and Wermter, S., "Real-world reinforcement learning for autonomous humanoid robot docking", *Robotics and Autonomous Systems,* Vol. 60, pp. 1400-1407, (2012).
27. Miljković, Z. Mitić, M. Lazarević, M. and Babić, B., "Neural network reinforcement learning for visual control of robot manipulators", *Expert Systems with Applications,* Vol. 40, pp. 1721-1736, (2013).
28. Kober, J., Bagnell, J.A. and Peters, J., "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research,* Vol. 32, pp. 1238-1274, (2013).
29. Fernandez-Gauna, B., Lopez-Guede, J.M. and Graña, M., "Transfer learning with partially constrained models: application to reinforcement learning of linked multicomponent robot system control", *Robotics and Autonomous Systems,* Vol. 61, pp. 694-703, (2013).
30. Fernandez-Gauna, B., Ansoategui, I. Etxeberria-Agiriano, I. and Graña, M., "Reinforcement learning of ball screw feed drive controllers", *Engineering Applications of Artificial Intelligence,* (2014).
31. Zarandi, M.H.F. Moosavi, S.V. and Zarinbal, M., "A fuzzy reinforcement learning algorithm for

inventory control in supply chains", *The International Journal of Advanced Manufacturing Technology,* Vol. 65, pp. 557-569, (2013).

32. Parbhoo, S., "A reinforcement learning design for HIV clinical trials", (2014).
33. Syafiie, S. Tadeo, F. and Martinez, E., "Model-free learning control of neutralization processes using reinforcement learning", *Engineering Applications of Artificial Intelligence,* Vol. 20, pp. 767-782, (2007).
34. Syafiie, S. Tadeo, F., Martinez, E. and Alvarez, T., "Model-free control based on reinforcement learning for a wastewater treatment problem", *Applied Soft Computing,* Vol. 11, pp. 73-82, (2011).
35. Zhao, Y., Zeng, D., Socinski, M.A. and Kosorok, M.R. "Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer", *Biometrics,* Vol. 67, pp. 1422-1433, (2011).
36. Farjadian, A.B. Yazdanpanah, M.J. and Shafai, B., "Application of Reinforcement Learning in Sliding Mode Control for Chattering Reduction", *Proceedings of the World Congress on Engineering*, (2013).
37. Khan, S.G., Herrmann, G., Lewis, F.L., Pipe, T. and Melhuish, C., "Reinforcement learning and optimal adaptive control: An overview and implementation examples", *Annual reviews in control,* Vol. 36 ,pp. 42-59, (2012).
38. Berenji, H., and Jamshidi, M., "Fuzzy reinforcement learning for System of Systems (SOS)," *IEEE International Conference on Fuzzy Systems, FUZZ 2011, June 27, 2011 - June 30, 2011*, Taipei, Taiwan, pp. 1689-1694, (2011).
39. Berenji, H.R., "A reinforcement learning—based architecture for fuzzy logic control", *International Journal of Approximate Reasoning,* Vol. 6, pp. 267-292, (1992).
40. Lee, C.-C. and Berenji, H., "An intelligent controller based on approximate reasoning and reinforcement learning", *Intelligent Control, 1989. Proceedings., IEEE International Symposium,* pp. 200-205, (1989).
41. Zadeh, L.A., "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic", *Fuzzy Sets and Systems,* Vol. 90, pp. 111-127, (1997).